基于局部显著特征与全局关系建模的阿尔茨海默综合症识别

陈炜东[1]，邢晓芬[1*]，陈帅琦[1]，范为铨[1]，丁万[2]
(1. 华南理工大学 电子与信息学院，广东广州 510641；
2. 深圳优必选科技有限公司，广东深圳 518000)

摘　要：由于阿尔茨海默综合症（Alzheimer's disease，AD）缺乏有效的药物治疗，对其早期的轻度认知障碍（Mild cognitive impairment, MCI）做出诊断并提前进行干预十分重要。近几年来，研究者们大都将注意力放在特征的提取上，采用传统的机器学习方式进行预测，效果并不理想。因此，本文同时从特征与模型的角度出发，尝试提高系统的识别率。在特征方面，本文提出采用清晰度特征和延音特征，同时结合不同参数提取的对数梅尔谱特征进行 AD 与 MCI 的识别。在模型方面，本文基于神经网络，首先突出长音频的局部显著特征，随后对音频的全局关系进行建模。在阿尔茨海默综合症识别竞赛中，本文所提出的方法在各个指标上比基准模型高 4%-5%，证明了该方法的有效性。

关键词：神经网络；语音识别；认知能力检测；阿尔茨海默综合症
中图分类号：TN912.34　　**文献标识码：**C

Locally salient features and global relations modeling for Alzheimer's disease detection

CHEN Weidong[1], XING Xiaofeng[1*], CHEN Shuaiqi[1], FAN Weiquan[1], DING Wan[2]
1. School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China;
2. UBTECH Robotics Corp, Shenzhen 518000, Chian

Abstract: The diagnosis and early intervention of mild cognitive impairment (MCI) is important due to the lack of effective treatments for Alzheimer's disease (AD). Recently, researchers mostly focus on feature extraction and use traditional machine learning methods for recognition, which have unsatisfied results. Therefore, this paper goes to improve the performance of system from both feature and model perspectives. In terms of feature, we propose articulation and prolongation features and combine them with log-Mel spectrum features extracted with different parameters for AD and MCI recognition. In terms of model, we use neural networks to highlight the locally salient features of long audio and then model the global relations. In the Alzheimer's disease recognition competition, the proposed model outperforms the baseline model by about 4%-5% in all criteria, which demonstrates the effectiveness of our proposed method.

Key words: neural network; speech recognition; cognitive decline detection; Alzheimer's disease

1　引言

阿尔兹海默症是一种神经系统疾病，其起病隐匿，常见于 70 岁以上老年人，故又称为老年痴呆症。阿尔兹海默症患者在生活中常表现为:记忆障碍，语言障碍，视觉空间障碍，计算障碍等。根据世界卫生组织

---

* 通讯作者

所发布的《2019 年全球卫生估计报告》，阿尔茨海默症及其他痴呆症位列 2000 年至 2019 年十大疾病死因之一。目前，全球约有 5000 万人患有阿尔兹海默症，平均不到 3 秒就有一个新发病例。中国的阿尔兹海默症患者接近 1500 万，人数居全球首位。随着人口老龄化的加剧，阿尔兹海默症也越来越引起人们重视。虽然阿尔兹海默症一旦确诊就无法治愈，但如果能在其发病的前期，即还处于 MCI 阶段时尽早发现，并尽早治疗，能够有效地延缓病情发展。因此，如何及时发现 MCI 阶段患者成为了研究的重点。

有研究认为，β-淀粉样蛋白和 tau 蛋白是诱导阿尔兹海默症产生的病因。β-淀粉样蛋白和 tau 蛋白一旦聚集，就会形成蛋白板块和神经元纠结。这些蛋白扩散到大脑的其他部分，导致大脑萎缩，其中最先影响的区域便是海马体。海马体对人的语义理解有关系，它负责将单词组成为句子。刚开始出现病情时，患者的语言处理能力下降，难以找到合适的词汇来表达，从而导致出现语言不流畅的情况。此时患者说话速度变慢，停顿时长变长，并且出现口吃现象。随着病情的进一步加重，患者的症状越加明显，说话逻辑变得混乱。这些特征的出现，让使用语音识别 MCI 与 AD 患者成为可能。

相比其他模态，基于语音的阿尔兹海默综合症的识别研究开展较晚，但也已经取得了一些成就。文献[1-3]提出了基于阿尔兹海默症的语音数据库，为推动这一领域的研究提供了便利。文献[4-6]等尝试将静态时长作为特征应用在语音识别阿尔兹海默症上，并证明了静态特征的有效性；文献[7]尝试将复杂度特征应用于阿尔兹海默症语音识别；文献[8]首次将词汇可用性理论应用在阿尔兹海默症识别，并认为阿尔兹海默症患者有自己的语料库。文献[9]使用了支持向量机（Support vector machine, SVM），朴素贝叶斯算法（Naive Bayers, NB）和 Adaboost 算法预测阿尔兹海默症。同时，也有一些将语音与文本结合的多模态研究[10-12]表明，多模态模型识别阿尔兹海默综合症的效果比单一模态效果更好。然而，目前大多数的研究都是基于英文环境下的阿尔茨海默症识别，鲜有针对中文群体的研究。同时，近期的研究大多集中在特征工程上，仍然使用传统的机器学习模型进行识别，效果并不理想。为了解决上述问题，本文同时从特征和模型研究的角度出发。在特征方面，本文提出使用清晰度特征与延音特征，并结合使用不同分辨率的对数梅尔谱特征同时进行阿尔茨海默综合症的识别。多种特征同时使用，能更加充分地挖掘出音频中的信息，提高识别效率。在模型方面，本文提出首先突出语音中的局部显著特征，随后再进行全局关系建模的方式，更加符合长音频的特性，提高建模效率。该方法在由江苏师范大学、清华大学和海天瑞声联合举办的阿尔茨海默综合症识别竞赛中取得了 84%的准确率，证明了该方法的有效性。

本文的结构安排如下：第二部分具体阐述多种特征的提取方式以及模型详细的建模步骤。第三部分描述实验的设置以及结果的比较与分析。最后一部分为结论。

2　本文提出的方法

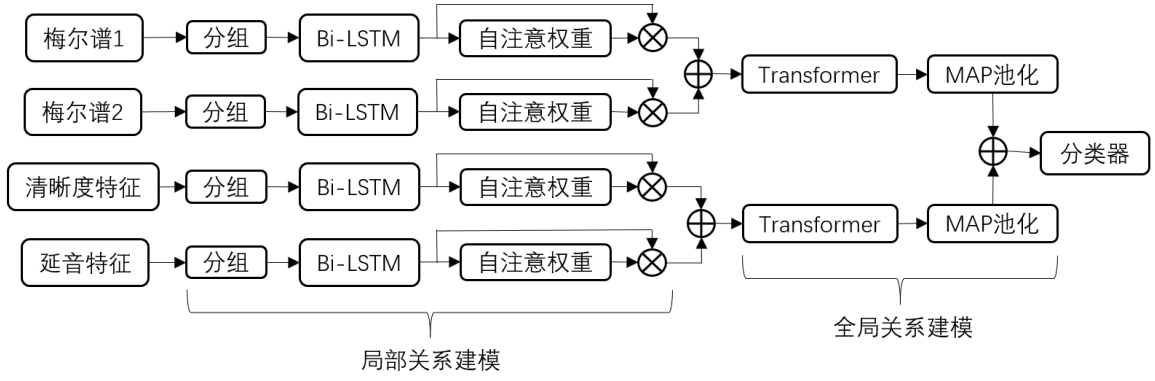

图 1　本文提出的阿尔茨海默综合症识别模型框架

Fig.1　The framework of the proposed AD recognition model

图 1 给出了本文提出的基于局部显著特征与全局关系建模的模型框架。其中，模型使用四种特征作为输入，分别是两种通过不同参数提取的对数梅尔谱特征、说话清晰度特征以及延音特征。局部关系的建模

用于突出局部显著特征，包括分组、双向长短时记忆网络[13]（Bi-directional long short-term memory, Bi-LSTM）以及一个计算自注意权重的操作。全局关系建模依靠 Transformer[14]与最大-平均池化（Max-average pooling, MAP）来实现。最后使用一个由两层全连接层构成的分类器来输出模型的最终预测结果。在接下来的小节里，本文将按照模型识别的流程，依次详细介绍输入特征的提取、局部关系建模以及全局关系建模方法，最后介绍本文使用的多路特征融合方法。

2.1 特征提取

2.1.1 对数梅尔谱特征

人耳听到的声音高低与实际的声音频率并不呈线性关系。梅尔谱基于此特性，在语谱图的基础上，将语谱图的频率轴由赫兹（Hz）刻度转换为梅尔刻度（Mel）。又因为声音存在掩蔽效应，在同一频率群中的声音，其能量会互相叠加。人耳基底膜则会对多个频率群中叠加后的能量做进一步的处理。因此，通过设计一个梅尔滤波器组，并将其作用在语谱图的能量谱上，即可得到梅尔谱特征。人耳对声强的感知也不是线性的，要使音量翻倍，声强大的地方要比声强小的地方增加更多的能量。因此，本文对提取的梅尔谱特征做对数运算，得到对数梅尔谱特征。

在提取梅尔谱特征时，需要进行短时傅里叶变换。在变换时，分帧使用的窗长越大，语谱图的频率分辨率越高，相应的时间分辨率则会降低。为了充分利用音频的频域与时域信息，本文分别采用 512 与 1024 的窗长，提取一段音频的两种不同分辨率的梅尔谱特征。其中，512 窗长提取的对数梅尔谱特征（梅尔谱 1）注重音频的时域特性，能清楚反映基频在时间上的变化规律等。1024 窗长提取的对数梅尔谱特征（梅尔谱 2）注重音频的频域特性，能清晰看出各共振峰的分布情况等。同时使用两种不同分辨率的对数梅尔谱特征进行预测，能更全面地使用音频的时域与频域信息，提高系统的性能。在提取对数梅尔谱特征时，帧移设为各自窗长的一半，梅尔滤波器组中滤波器数量为各自窗长的四分之一。

2.1.2 说话清晰度特征与延音特征

在识别阿尔茨海默综合症的过程中，说话人的说话清晰度以及说话结束时的拉长所造成的延音特性，具有明显的标识性，对识别能起到有效的帮助。因此，本文提取音频中的说话清晰度特征与延音特征。其中，说话清晰度关注说话者说话开始瞬间的信息，延音特征关注说话者说话结束瞬间的信息。

要提取说话清晰度与延音特征，需要知道音频中说话开始与结束时刻的时间戳信息。本文首先使用 praat 工具提取音频中的基频，通过语音的基频特性，将基频由 0 转至大于 0 的瞬间视为说话开始的时刻，由大于 0 转至 0 的瞬间视为说话结束的时刻。对于每一个时刻，取它左右各 40 毫秒的音频形成一个小段。对于每一个小段音频进行傅里叶变换，其中帧长设为 40 毫秒，帧移为 20 毫秒，得到每一个音频小段的语谱图。将 22 个 Bark 滤波器作用在其语谱图上进行滤波，得到 22 个 Bark 子带能量作为该音频小段的特征。另外，本文继续提取每一小段音频的 12 维梅尔频率倒谱系数（Mel-frequency cepstral coefficients, MFCC）及其 12 维的一阶差分与 12 维的二阶差分特征，共得 36 维的 MFCC 特征。结合 22 维的 Bark 子带能量特征，共 58 维。将所有代表说话开始的小段音频所提取出来的特征按照时间顺序拼接在一起，形成该音频的说话清晰度特征。同理，将所有代表说话结束的小段提取出来的特征拼接在一起，构成延音特征。

2.2 基于 Bi-LSTM 的局部关系建模

2.2.1 基于特征分组的 Bi-LSTM 模型

本系统使用的输入音频长度约为一分钟，属于长语音的分类模型。如果直接对长语音进行建模，输入的数据量大，模型的参数量大，导致模型学习困难。因此，在对长语音进行建模的过程中，本文先对输入的特征在时间序列上进行分组，每一个小组的特征代表长音频中的某一个局部的特征。由此可将一个长音频分解成多个局部的组合。最后让 Bi-LSTM 对每一个小组特征进行学习，取 Bi-LSTM 最后一个时刻的输出作为该局部的特征表示。通过特征的分组与 Bi-LSTM 的学习，可以让系统对长语音的局部关系进行充分建模，同时大大降低了模型的复杂度与计算量，减轻模型的过拟合现象。

2.2.2 自注意机制

虽然特征分组将长音频分解成多个局部的组合，并通过 Bi-LSTM 学习到每一个组的特征表达。但是，

并不是每一个局部小组的重要性都是一样的。比如，某些出现明显阿尔茨海默特性的说话片段出现在某一个局部，那么该局部的小组是比较重要的。当长音频中的某些局部没有出现说话声音，是空白的片段，那么这些局部的小组是不重要的。因此，本文提出一种自注意机制，旨在让模型学会自动根据不同小组的重要程度，从而给不同的小组分配不同的权重。

假设给定输入特征$X \in R^{t \times d_{in}}$，其中t为时间序列长度，$d_{in}$为输入特征维度。通过特征分组、Bi-LSTM学习后得到$X_g \in R^{g \times d}$，其中 g 为人为设定的分组组数，d 为 Bi-LSTM 的输出维度。本文使用的自注意机制的实现过程可描述为如下：

$$X_g = \mathrm{BiLSTM}(\mathrm{group}(X)) \tag{1}$$

$$e = X_g \times X_g^T \tag{2}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{j=1}^{g} exp(e_{ij})} \tag{3}$$

$$X_a = \alpha \times X_g \tag{4}$$

上式中，group(·)表示分组操作，T表示矩阵的转置，α为经过 softmax 函数规整为 0 到 1 之间的自注意权重矩阵，$X_a \in R^{g \times d}$为自注意机制的输出。通过自注意机制让模型自动判断不同局部特征之间的重要性，并赋予不同的权重，从而突出显著的局部特征。

2.3 基于 Transformer 的全局关系建模

通过分组与 Bi-LSTM 结构，系统对长语音的局部特征进行了充分的建模，并通过自注意机制突出了局部显著的特征。然而，系统还缺少对整段音频的全局关系学习。因此，本文利用 Transformer 对长序列优异的建模能力，来提取整段音频的全局特征。

2.3.1 Transformer 结构

Transformer 抛弃了传统的循环神经网络结构，将输入分为了 Q，K，V 三部分，分别代表查询，键，值向量。为了保持序列中的位置信息，Transformer 对输入的序列进行位置编码。包含 ReLU 激活函数[15]的前馈网络为 Transformer 提供非线性变换。多头注意力机制是 Transformer 的核心部分，它的主要思想是通过Q和K的相乘，从而在$\boldsymbol{V}$中找到每个查询向量最相关的部分，并产生一个权重矩阵W。通过将W和V相乘，得到注意力输出$attn$。其计算过程可描述如下：

$$W = softmax\left(\frac{QK^T}{\sqrt{d_Q}}\right) \tag{5}$$

$$attn = W \times V \tag{6}$$

其中，d_Q是查询向量的维度。对于多头注意力机制，将每一个头的注意力输出拼接在一起，然后输入进一个线性变换层，得到最终的输出。假设$attn_i$是第i个头的注意力输出，总共的头数是M，多头注意力机制的计算定义如下：

$$A = concat(attn_1, attn_2, attn_3 \ldots attn_M)W_o \tag{7}$$

其中，W_o是线性变换层的可学习参数，A是多头注意力机制最终的输出。Transformer 的具体实现与计算细节可参考原文[14]，此处不进行赘述。

2.3.2 最大-平均池化

在以往的深度学习模型中，常用的池化方式有最大池化和平均池化两种。然而，对于语音输入的情况而言，最大池化的结果容易受到噪声的影响，表现不稳定。而且语音中的有效识别信息往往集中在一个小的区域之内，长音频中表现出阿尔茨海默综合症特性的语音也只占整体音频的一小部分。如果使用平均池化，容易把这些有效信息给均分、削弱，不利于模型的学习。因此，本文提出一种新型的池化方式，称为最大-平均池化（MAP），通过只选择响应较大的特征进行平均池化，既保留有效性息，同时减轻噪声对结果的干扰，更加适合语音信号的特性。MAP 的计算过程描述如下：

$$X_{Trans} = Trans(X_a) \tag{7}$$

$$S = sum(X_{Trans}) \tag{8}$$

$$threshold = topk(\mathrm{S}) \tag{9}$$

$$S_{mask} = \begin{cases} 0, & S_i < threshold \\ 1, & S_i \geq threshold \end{cases} \tag{10}$$

$$S_{sum} = sum(S_{mask} * X_{Trans}) \tag{11}$$

其中，$Trans(\cdot)$代表 Transformer 结构，sum(·)代表求和，topk(·)代表求解第 k 大值。$X_{Trans} \in R^{g \times d}$，$\mathrm{S} \in R^{g \times 1}$，$S_{mask} \in R^{g \times 1}$，$*$代表矩阵的对应位置相乘，$S_{sum} \in R^{1 \times d}$为最大-平均池化的输出。

2.4 特征融合

特征融合的方式一般分为两种：早期融合（Early fusion）与晚期融合（Late fusion）。早期融合指在特征学习的早期过程中进行融合，能够更有效地学习特征之间的交互信息，融合的程度更深。晚期融合一般指发生在决策层或是靠近决策层的融合，操作简单，融合程度较浅。本文先对相同类型的特征进行早期融合，随后将不同类型的特征进行晚期融合。具体如图 1 所示，两种对数梅尔谱特征经过各自的 Bi-LSTM、自注意权重加权后相加，进行早期融合。同理，清晰度特征与延音特征也在相同的位置进行早期融合。四路的特征经过早期融合后形成两路特征。这两路特征经过各自的 Transformer、MAP 池化后进行相加，完成晚期融合，合并成为一个融合特征。将最后的融合特征输入分类器，得到最终的预测结果。

3 仿真实验

3.1 实验数据及设置

本文使用的数据是由比赛方提供的长语音音频。音频分别从 AD、MCI 和正常人（Healthy control, HC）中采集，内容为图片描述和流畅性命名等。音频最长时长为 1 分钟，共计 280 条训练数据，119 条测试数据。由于官方并没有提供验证集的划分方法，本文使用了两种交叉验证策略。第一种策略是对 280 条训练数据进行随机五折交叉验证（Corss-validation, CV）划分。然而，随机划分会使同一说话人的音频同时出现在训练集和验证集，导致说话人信息泄露。因此，本文使用的第二种策略，是根据说话人来划分训练集与验证集，确保同一个说话人的音频只出现在其中一个集内，形成说话人独立的三折交叉验证划分。模型性能用所有折的平均性能表示。综合考虑模型在两种交叉验证策略上的性能表现，挑选出最优的参数设置，并用全部的训练数据进行训练，得到最终模型。模型在测数数据上进行测试，得到竞赛的测试结果。

本文基于 PyTroch 搭建识别模型。Bi-LSTM 层数为 1，隐藏层维度为 128。Transformer 层数为 2，随机丢弃[16]（Dropout）概率设为 0.5 以缓解模型的过拟合。MAP 池化中 k 设置为 50%，即选择输入中前 50%的较大值进行平均池化。窗长为 512 提取出来的对数梅尔谱特征的特征维度是 128，窗长为 1024 提取出来的特征维度则是 256。清晰度特征与延音特征的特征维度是 58。两个对数梅尔谱特征在时间序列上平均分为 180 组，清晰度特征与延音特征在时间序列上平均分为 50 组。本文采用反向传播算法更新模型参数。

为了提高模型的泛化能力，本文尝试了两种数据增强方式：mixup[17]、copypaste[18]。数据增强方法的详细介绍可参考原文[17, 18]，此处不进行赘述。识别率（Accuracy）、查准率（Precision）、召回率（Recall rate）和 F1 值将作为模型的评价指标。

3.2 实验结果与分析

为了验证多特征输入对模型的好处以及验证清晰度特征与延音特征的有效性，本文首先在说话人独立的三折交叉验证设置下进行了消融实验，结果如表 1 所示。对比表 1 中的前三行，可以看出，同时使用不同分辨率的对数梅尔谱特征比单独使用一种分辨率的对数梅尔谱特征都要好。对比表中第三行与第四行可以看出，加上清晰度特征与延音特征后能让模型的性能进一步提升，证明了清晰度与延音特征的有效性。

表 1　在说话人独立的三折交叉验证下的消融实验

Tab.1　Ablation studise in speaker-independent 3-folds cross-validation

特征	Accuracy	Precision	Recall rate	F1
梅尔谱 1	0.710	0.728	0.714	0.712

梅尔谱 2	0.715	0.736	0.724	0.721
梅尔谱 1+梅尔谱 2	0.731	0.758	0.731	0.728
梅尔谱 1+梅尔谱 2+清晰度特征+延音特征	0.743	0.766	0.740	0.731

将本文提出的模型简称为 m1，结构如图 1 所示。在 m1 的基础上进行修改，使用两个对数梅尔谱特征与延音特征，并去掉早期融合，仅通过后期融合将三条通路的信息进行合并，形成模型 m2。mixup 与 copypaste 数据增强方式分别应用在模型 m1 与 m2 上。比如，m1+mixup 表示使用 mixup 数据增强方式训练模型 m1。为了进一步提高系统的性能，同时提高模型的稳定性与鲁棒性，本文尝试了多种模型集成方案 Ensemble。其中，效果最好的集成方案是 m1+mixup 与 m2+copypaste 的集成，本文仅展示在该集成方案下的性能表现。表 2 和表 3 分别给出了在不同交叉验证策略下的识别结果对比。

从表 2 和表 3 中可以看出，无论是五折随机交叉验证还是说话人独立的三折交叉验证，本文提出的方法，都比官方提供的基准 Baseline 要好。尤其是在说话人独立的三折交叉验证下，本文提出的方法在各指标上都比 Baseline 高 20%以上。而在五折随机交叉验证下，模型的性能普遍较高，本文提出的方法在各个指标上比 Baseline 高 5%左右。

表 2　五折随机交叉验证结果对比

Tab.2　Comparision results in randomly split 5-folds cross-validation

模型	Accuracy	Precision	Recall rate	F1
Baseline (SVM)	0.846	0.856	0.846	0.845
Ours (m1)	**0.903**	**0.911**	**0.905**	**0.904**
Ours (m1+mixup)	0.893	0.901	0.896	0.891
Ours (m1+copypaste)	0.875	0.886	0.880	0.875
Ours (m2)	0.875	0.877	0.878	0.873
Ours (m2+mixup)	0.875	0.881	0.879	0.875
Ours (m2+copypaste)	0.879	0.888	0.882	0.879
Ours (Ensemble)	0.875	0.885	0.879	0.875

表 3　说话人独立的三折交叉验证结果对比

Tab.3　Comparision results in speaker-independent 3-folds cross-validation

模型	Accuracy	Precision	Recall rate	F1
Baseline (SVM)	0.598	0.589	0.576	0.518
Ours (m1)	0.743	0.766	0.740	0.731
Ours (m1+mixup)	0.747	0.779	0.740	0.729
Ours (m1+copypaste)	0.749	0.764	0.745	0.736
Ours (m2)	0.813	0.822	0.806	0.814
Ours (m2+mixup)	**0.820**	**0.824**	**0.819**	**0.818**
Ours (m2+copypaste)	0.801	0.818	0.800	0.796
Ours (Ensemble)	0.792	0.807	0.790	0.781

最终的测试结果如表 4 所示。本文挑选了两个模型进行测试，分别是 m1+mixup 与 Ensemble。测试结果显示，本文提出的模型在各个指标上比 Baseline 高 4%-5%，证明了本文所提出的模型的有效性。

表 4　测试集结果对比

Tab.4　Comparision results in test dataset

模型	Accuracy	Precision	Recall rate	F1
Baseline (SVM)	0.798	0.799	0.785	0.786
Ours (m1+mixup)	0.824	0.821	0.808	0.809
Ours (Ensemble)	**0.840**	**0.847**	**0.841**	**0.838**

4 结论

针对阿尔茨海默综合症，本文提出使用说话清晰度与延音特征，并结合不同分辨率的对数梅尔谱特征进行识别。同时，针对长音频输入，本文提出先分组并通过 Bi-LSTM 对音频的局部信息进行建模，减少系统的参数量，随后通过 Transformer 学习全局的依赖关系，得到整段音频的特征表达。针对语音信号的特性，本文提出一种自注意机制与最大-平均池化，帮助模型更有效地学习音频中的信息。在阿尔茨海默综合症识别竞赛中，本文所提出的模型取得了准确率为 84%的性能表现。在未来的工作中，作者希望将此系统与语音识别系统结合，提取音频中的文字信息，采用多模态的方式进行阿尔茨海默综合症的识别，进一步提升模型准确率和鲁棒性。

参考文献

[1] MacWhinney B, Fromm D, Forbes M, et al. Aphasiabank: Methods for studying discourse[J]. Aphasiology, 2011, vol. 25, no. 11, pp. 1286–1307.

[2] Acosta-Baena N, Sepulveda-Falla D, Lopera-Gómez C M, et al. Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: a retrospective cohort study[J]. The Lancet Neurology, 2011, vol. 10, no. 3, pp. 213–220.

[3] Luz S, Haider F, de la Fuente S, et al. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge[C]//Proc. Interspeech. 2020: 2172–2176.

[4] Yuan J, Cai X, Church K. Pause-Encoded Language Models for Recognition of Alzheimer's Disease and Emotion[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 7293-7297.

[5] Rohanian M, Hough J, Purver M. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs[J]. arXiv preprint arXiv: 2106.15684, 2021.

[6] Syed Z S, Syed M S S, Lech M, et al. Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech[C]//Proc. Interspeech. 2021: 3815-3819.

[7] Pérez-Toro P A, Bayerl S P, Arias-Vergara T, et al. Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge[C]//Proc. Interspeech. 2021: 3785-3789.

[8] Villatoro-Tello E, Dubagunta P, Fritsch J, et al. Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition[C]//Proc. Interspeech. 2021: 1927-1931.

[9] Rudzicz F, Chan Currie L, Danks A, et al. Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease[C]//Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility. 2014: 241-242.

[10] Wang N, Cao Y, Hao S, et al. Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data[C]//Proc. Interspeech. 2021: 3835-3839.

[11] Pan Y, Mirheidari B, Harris J M, et al. Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech[C]//Proc. Interspeech. 2021: 3810-3814.

[12] Li J, Yu J, Ye Z, et al. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6423-6427.

[13] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Proc. Interspeech. 2014: 338-342.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[15] Agarap A F. Deep learning using rectified linear units (relu)[J]. arXiv preprint arXiv: 1803.08375, 2018.

[16] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[17] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv: 1710.09412, 2017.

[18] Pappagari R, Villalba J, Żelasko P, et al. CopyPaste: An Augmentation Method for Speech Emotion Recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6324-6328.

作者简介

陈炜东 男，1999，广东。华南理工大学，在读硕士生，研究方向为语音情感计算。E-mail: eewdchen@mail.scut.edu.cn

邢晓芬 女，1979，江苏。华南理工大学，副教授，博士，研究方向为语音情感计算。E-mail: xfxing@scut.edu.cn

陈帅琦 男，1999，广东。华南理工大学，在读硕士生，研究方向为语音情感计算。E-mail：202121013939@mail.scut.edu.cn

范为铨 男，1996，广东。华南理工大学，在读博士生，研究方向为语音情感计算。E-mail：weiquan.fan96@gmail.com

丁万 男，1983，武汉。优必选软件技术（深圳）有限公司，专家工程师，博士，研究方向为多模态情感识别与多模态情感合成。E-mail: wan.ding@ubtrobot.com

创新点说明

在特征方面，本文提出使用说话清晰度与延音特征，并结合不同分辨率的对数梅尔谱特征，充分地挖掘出音频中所包含的信息。并且本文通过早期融合与晚期融合相结合的方式，有效使用所有的输入特征。在模型方面，针对长音频的特性，本文先对局部信息进行建模，并突出局部显著特征，随后再对全局关系进行建模。该建模方式能有效降低模型的参数量，同时突出音频中的关键部分。针对语音信号，本文还提出一种自注意机制与最大-平均池化方法，帮助模型更有效地学习语音信号中的信息。